Action Recognition from Arbitrary Views using 3D Exemplars

Daniel Weinland, Edmond Boyer, Rémi Ronfard

Action Recognition from Arbitrary Views using 3D Exemplars

Daniel Weinland*
LJK - INRIA
Grenoble, France
weinland@inrialpes.fr

Edmond Boyer
LJK - INRIA
Grenoble, France
eboyer@inrialpes.fr

Remi Ronfard
Artificialife Inc.
Montreal, Canada
remir@artificialife.com

Abstract

In this paper, we address the problem of learning compact, view-independent, realistic 3D models of human actions recorded with multiple cameras, for the purpose of recognizing those same actions from a single or few cameras, without prior knowledge about the relative orientations between the cameras and the subjects. To this aim, we propose a new framework where we model actions using three dimensional occupancy grids, built from multiple viewpoints, in an exemplar-based HMM. The novelty is, that a 3D reconstruction is not required during the recognition phase, instead learned 3D exemplars are used to produce 2D image information that is compared to the observations. Parameters that describe image projections are added as latent variables in the recognition process. In addition, the temporal Markov dependency applied to view parameters allows them to evolve during recognition as with a smoothly moving camera. The effectiveness of the framework is demonstrated with experiments on real datasets and with challenging recognition scenarios.

1. Introduction

We consider the problem of recognizing actions using *a priori* unknown camera configurations. Action recognition has received considerable attention over the past decades, as a result of the growing interest for automatic and advanced scene interpretations shown in several applications domains, *e.g.* video-surveillance or human machine interactions. In this field, two main directions have been followed. *Model based approaches*, *e.g.* [6, 20] assume a known parametric model, typically a kinematic model, and represent actions in a joint or parameter space. Unfortunately, recovering the parameters, *e.g.* the pose, of the model appears to be a difficult intermediate task without the help of landmarks.

*D. Weinland is supported by a grant from the European Community under the EST Marie-Curie Project Visitor.

In contrast, *template based* or *holistic* approaches, *e.g.* [3, 7, 2, 19], do not use such an intermediate representation and directly model actions using image information, silhouettes or optical flow for instance. Action templates are then spatio-temporal shapes either in a three-dimensional space, when a single camera is considered, or in a four dimensional space when multiple calibrated cameras are considered. In both cases, action recognition is achieved by comparing a motion template, built from observations, with learned models of the same type. This limits recognition to situations where observed and learned models are obtained using similar camera configurations.

In this work, we propose an approach that takes advantage of the template based methods but that does not constrain camera configurations during recognition. Instead, actions can be observed with any camera configuration, from single to multiple cameras, and from any viewpoint. Our main motivation is to be able to cope with unknown recognition scenarios without learning multiple and specific databases. This has particularly clear applications in video-surveillance where actions are often observed from a single and arbitrary viewpoint.

To this purpose, we propose an exemplar-based hidden Markov model (HMM) inspired by the works of Frey and Jojic [9] and Toyama and Blake [18]. This model accounts for dependencies between three dimensional exemplars, *i.e.* representative pose instances, and image cues, this over time sequences. Inference is then used to identify the action sequence that best explains the image observations. In particular, a nice feature is that observations from any calibrated view can be incorporated. In addition, explicitly modeling the transformation between exemplars and image cues allows such transformation to change over time during recognition.

The paper proceeds as follows. In Section 2 we review the state of the art in view-independent action recognition. In Section 3 we present an overview of the proposed approach. Details on the exemplar-based HMM design are given in Section 4. In Section 5 the exemplar selection and the model learning are explained. Section 6 details recogni-

tion. Experiments using a challenging dataset of 11 actions are presented in Section 7.

2. Related Work

In order to allow actions to be learned and recognized using different camera configurations, action descriptions must exhibit some view invariance. Campbell [5] describes 3D hand and head trajectories using view invariant coordinate representations. Fundamental matrices can also be used to compare 2D action representations from different views, as joint trajectories in [16, 20] or silhouettes in [17]. To achieve similar comparisons, Parameswaran and Chellappa [14] use projective invariants of coplanar landmark points on a human body. In a previous work [19] we compare 3D action representations based on visual hulls and propose invariant Fourier-descriptors that are computed from multiple-view reconstructions. These approaches have focused on representations in which view dependent information is removed, often at the cost of an impoverished action model and without adding full flexibility in camera configurations. This motivates the search for another solution.

In a different context, Frey and Jojic [9] show how to account for view transformations in a dynamic probabilistic model. In the same spirit, Toyama and Blake [18] extend the idea for tracking with powerful image distances, and Elgammal *et al*. [8] propose a nonparametric mixture extension that, however, applies to view-dependent action recognition. Our approach builds on a similar model and incorporates geometric transformations into the probabilistic modeling of an action.

It is worth to mention also the work of Brand[4] that uses HMMs and a direct mapping between a three dimensional joint space and silhouette observations for pose estimation. It shares some similarities with our approach since we also use HMMs to model temporal sequences of exemplars.

A very recent and interesting work is that of Lv and Nevatia [12]. Developed in parallel to our method, it shares the idea of projecting a set of learned 3D exemplars/key-poses into 2D to infer actions from arbitrary view. However we use a probabilistic model instead of the deterministic linked action graph introduced in [12], allowing therefore to naturally handle uncertainties inherent to actions performed by different people and with different styles.

3. Overview

We model an action as a sequence over a set of key-poses, the exemplars. Figure 1 shows two examples of observation sequences and the corresponding best matching exemplar sequences computed with our model.

Exemplars are represented in 3D as visual hulls that have been computed using a system of 5 calibrated cameras. The model does thus not rely on motion capture data, which is generally difficult to obtain.

The observation sequence comes in this example from a single camera and is represented trough silhouettes obtained from background subtraction. To match observation and exemplars, the visual hulls are projected into 2D and a match between the resulting silhouettes is computed. The recognition phase thus generates 2D from 3D and never has to infer 3D from a single view observation.

Modeling actions and views The matching between model and observation is represented in a probabilistic framework (Section 4). Consequently, and crucially, that neither the best matching exemplar sequence, nor the exact projection parameters need to be known. Instead a probability of all potential exemplar sequence and projection is computed. Using the classical HMM algorithms [15], such a probability can be efficiently computed under the following conditions: First, we use a small set of exemplars that is shared by all models. As we show in Section 5.1, a small set of exemplars is sufficient to describe a large variety of actions, if the exemplars are discriminative with respect to these actions. Second, we make a few reasonable assumptions on the parameters of the projective transformation, *i.e*. the camera calibration and position of a person can be robustly observed during recognition and only the orientation of a person around the vertical axis is unknown.

Exemplar selection and model learning Learning an action model consists of two steps: A set of exemplars is selected and shared by all actions models (Section 5.1); probabilities over these exemplars are learned individually for each action (Section 5.2).

When selecting the exemplars, we are interested in finding the subset of poses from the training sequences, that bests discriminates actions. To this purpose, we present in Section 5.1 a novel solution based on a method for feature subset selection, a *wrapper* [11].

Given a set of exemplars, the action specific probabilities are estimated using standard probability estimation techniques for HMMs, as described in Section 5.2. Interestingly, the learning of dynamics over a set of selected 3D exemplars can be performed either on 3D sequences of aligned visually hulls (Section 5.2.1), thus under ideal conditions, or simply from single view observations (Section 5.2.2). Hence 3D information is not mandatory for that step.

Classification Classification is performed using standard HMM algorithms, as described in Section 6.

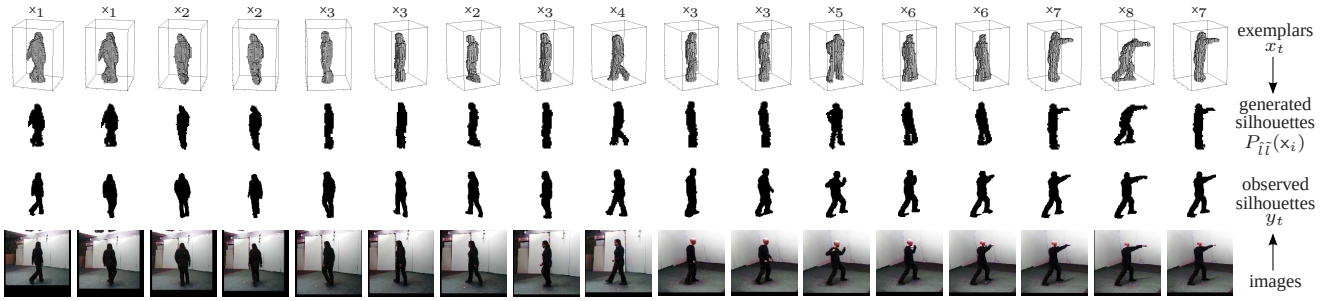

Figure 1. 2D observation sequences y_t ("walk in cycle" and "punch"), observed from different viewpoints and with unknown orientation of the persons, are explained trough 3D action models. The best matching exemplar sequence x_t and the best matching 2D projection $P_{\hat{l}\tilde{l}}(\mathsf{x}_i)$, as generated by the models, are displayed. Both models share a small set of exemplars (labeled on top).

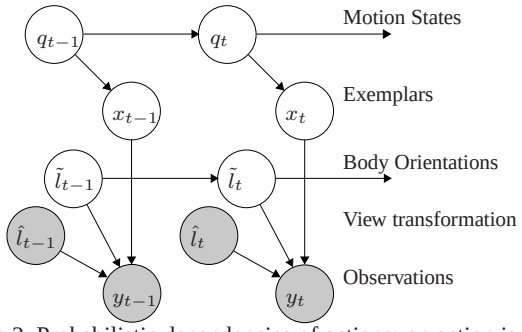

Figure 2. Probabilistic dependencies of actions: an action is modeled as a hidden state sequence Q, *e.g.* a motion sequence in a pose space. At each time step t, a 3D exemplar x_t, *i.e.* a visual hull, is drawn from the motion sequence Q. Observations y_t, *i.e.* silhouettes, result then from a geometric transformation of exemplars that is defined by 2 sets of parameters $\hat{l}$ and $\tilde{l}$. $\hat{l}$ are observed parameters, *e.g.* camera parameters determined in a preliminary step, and $\tilde{l}$ are latent parameters, *e.g.* body orientation determined during recognition. Shaded nodes in the graph correspond to observed variables.

4. Probabilistic Model of Actions and Views

Our representation for human action is a product of two independent random processes, one for the orientation of the subject relative to the camera, and the other for the view-independent, body-centered poses taken by the performer during the various stages of the action. The two processes are modeled in an exemplar based Markov model, shown in Figure 2, in the spirit of [9] and [18].

Hidden Motion States Dynamics in exemplar space are represented by a discrete N-state latent variable q that follows a first order Markov chain over time. Thus: $p(q_t|q_{t-1},\ldots,q_1) = p(q_t|q_{t-1})$, with $t \in [1\ldots T]$, and with the prior $p(q_1)$ at time $t = 1$. Though generally hidden, q can intuitively be interpreted as a quantization of the joint motion space into action-characteristic configurations.

Exemplars At each time t, a three dimensional body template x_t is drawn from $p(x_t|q_t)$. A crucial remark here is that these templates do not result from body models and joint configurations but are instead represented by a set of M exemplars: $\mathsf{X} = \{\mathsf{x}_{i\in[1\ldots M]}\}$, learned from three dimensional training sequences.

Note here that $p(x_t = \mathsf{x}_i|q_t)$ models the non-deterministic dependencies between motion states and body configuration. Thus motion states q are not deterministically linked to exemplars as in [12, 18], allowing therefore a single motion state q to be represented with different exemplars, to account for different body proportions, style, or clothes.

View Transformation and Observation To ensure independence with respect to the view projection onto the image plane: $P_{\hat{l}\tilde{l}}(\mathsf{x}) = \hat{P}[R_\theta, u]\mathsf{x}$, we condition observations y on parameters that represent this transformation. We differentiate view transformation parameters $\{\hat{l}_t\}$ that can be robustly observed (*i.e.* the camera matrix $\hat{P}$ and position u), and body pose parameters $\{\tilde{l}_t\}$ that are latent (*i.e.* the orientation around the vertical axis θ).

The resulting density $p(y_t|x_t, \hat{l}_t, \tilde{l}_t)$ is represented in form of a kernel function centered on the transformed exemplars $P_{\hat{l}\tilde{l}}(\mathsf{x}_i)$:

$$p(y_t|x_t = \mathsf{x}_i, \hat{l}_t, \tilde{l}_t) \propto \frac{1}{Z} \exp\big(-d(y_t, P_{\hat{l}\tilde{l}}(\mathsf{x}_i))/\sigma^2\big), \quad (1)$$

where d is a distance function between between the resulting silhouettes, *e.g.* the Euclidean distance (i.e. the number of pixels which are different), or a more specialized distance such as the chamfer distance [10]. (Note that both were giving similar results in our experiments.)

The temporal evolution of the latent transformation variables is modeled as a Markov process with transitions probabilities $p(\tilde{l}_t|\tilde{l}_{t-1})$, and a prior $p(\tilde{l}_1)$. This is equivalent to a temporal filtering of the transformation parameters where, interestingly, various assumptions could be made on the dynamic of these parameters: a static model or an autoregres-

sive model, or even a model taking into account dependencies between an action and view changes.

In our implementation all variables $\{\tilde{l}, \hat{l}\}$ are discretized. For instance, the orientation θ is discretized into L equally spaced angles within $[0, 2\pi]$ and u is discretized into a set of discrete positions. The temporal evolution of θ is modeled using a von Mises distribution: $p(\theta_t|\theta_{t-1}) \propto \exp(\kappa \cos(\theta_t - \theta_{t-1}))$, that can be seen as the circular equivalent of a normal distribution, and a uniform prior $p(\theta_1)$.

5. Learning

We learn separate action models λ_c for each action class $c \in \{1, \ldots, C\}$. A sequence of observations $Y = \{y_1, \ldots, y_T\}$ is then classified with respect to the maximum a posteriori (MAP) estimate:

$$g(Y) = \arg\max_c p(Y|\lambda_c)p(\lambda_c). \quad (2)$$

The set λ_c is composed of the probability transition matrices $p(q_t|q_{t-1}, c)$, $p(q_1|c)$ and $p(x_t|q_t, c)$, which are specific to the action c, as they represent the action's dynamics. In contrast, the observation probabilities $p(y_t|x_t, \hat{l}_t, \tilde{l}_t)$ are tied between classes, meaning that all actions $\{c = 1..C\}$ share a common exemplar set, *i.e.* $\mathsf{X}_c = \mathsf{X}$, and a unique variance $\sigma_c^2 = \sigma^2$. In the context of HMMs, such an architecture is known as a *tied-mixture* or *semi-continuous* HMM[1]. This architecture is particularly well adapted to action recognition since different actions naturally share similar poses. For example, many actions share a neutral rest position and some actions only differ by the sequential order of poses that composed them. In addition, sharing parameters dramatically reduces complexity during recognition, when every exemplar must be projected with respect to numerous latent orientations.

Learning consists then in two main operations: selecting the exemplar set that is shared by all models; learning the action specific probabilities. As we will see in the following, the two operations are tightly coupled. Selection uses learning to evaluate the discriminant quality of an candidate exemplar set, and learning probabilities relies on a selected set of exemplars. Both operations are detailed below.

5.1. Exemplar Selection

Identifying discriminative exemplars is an essential step of the learning process. Previous works use motion energy minima and maxima [12, 13], or k-means clustering (adapted to return exemplars) [18] to this end. However, there is no apparent relationship between such criteria and the action discriminant quality of the selected exemplars. In particular for the adapted k-means clustering [18] we observed experimentally, that clusters tend to consist of different poses performed by similar actors rather than similar poses performed by different actors. Consequently, selecting exemplars as poses with minimum within-cluster distance often leads to neutral and therefore non-discriminative poses.

In light of this, we propose a novel approach for exemplar selection, to better link the discriminant quality of exemplars and the selection. We therefore use a wrapper [11], a technique for discriminant feature subset selection. The idea behind a wrapper is to use the trained classifier (2) itself to evaluate how discriminative a candidate set of exemplars is. Thus a wrapper performs a greedy search over the full set of exemplars, where in each iteration classifiers are learned and evaluated for each possible subset considered.

The wrapper method we use is called "forward selection" [11], and proceeds as follows: Let $\mathcal{Y}$ denote a set of 3D visual hulls. Assume training sequences and test sequences for all actions $c \in \{1, \ldots, C\}$ are given.

1. Set $\mathsf{X} = \emptyset$.

2. Find $y^* \in \{\mathcal{Y} \setminus \mathsf{X}\}$, where a classifier g (trained on all actions) using exemplar set $\{\mathsf{X} \cup y^*\}$ has best recognition performance on the test-set. Add y^* to X.

3. Repeat step 2 until M visual hulls from $\mathcal{Y}$ have been added to X.

Note that the above procedure can only work when the exemplar set is shared by all action models. The selection thus starts by training a classifier for each singleton exemplar. The exemplar for which the classifier has best evaluation performance is selected, and the procedure is repeated for couples of exemplars, triples, *etc*., until M exemplars have been selected. Note that training and evaluation of the classifier can be performed in 3D or 2D, as detailed in Section 5.2. In case that the training sequences are 3D, $\mathcal{Y}$ can simply be the training-set.

The approach is illustrated in Figures 3 and 4 where exemplars and the associated classification rates are shown. Figure 3 shows that the selected poses naturally represent *key-frames* or characteristic frames of an action.

5.2. Learning Dynamics

Given a set of exemplars, the action parameters $\lambda_{c\in\{1,\ldots,C\}}$: probabilities $p(q_t|q_{t-1}, c)$, $p(q_1|c)$ and $p(x_t|q_t, c)$, can be learned. Various strategies can be considered for that purpose. In the following, we sketch 2 of them: learning from 3D observations (sequences of visual hulls), and learning from 2D observations (image sequences). Note that in both cases, motion is learned in 3D over the set of 3D exemplars, obtained as described in section 5.1.

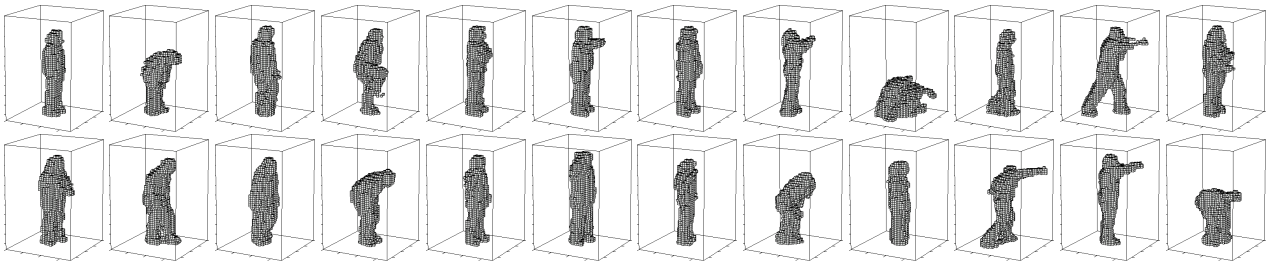

Figure 3. Selected exemplars: first 24 discriminative exemplars as returned by the forward selection. The dataset is composed of 11 actions performed by 10 actors. Recognition rates are shown in Figure 4.

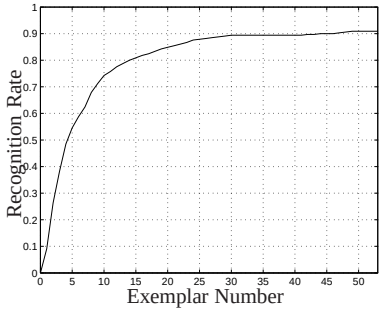

Figure 4. Recognition rate vs. number of selected exemplars.

5.2.1 Learning from 3D Observations

In this training scenario, several calibrated viewpoints are available, leading therefore to 3D visual hull sequences, and all actions are performed with the same orientation. In that case, motion dynamics are learned independently from any viewing transformation, thus $p(y_t|x_t, \hat{l}_t, \tilde{l}_t) = p(y_t|x_t)$ with y being 3D. Transformation parameters appear later during the recognition phase where both dynamics and viewing process are joined into a single model.

Each model λ_c is learned through a forward-backward algorithm that is similar to the standard algorithm for Gaussian mixture HMMs [15], except that the kernel parameters, that correspond to mean and variance of the Gaussians (*i.e.* X and σ), are not updated. Note that a similar forward-backward algorithm was already proposed in the context of exemplar based HMMs [8].

5.2.2 Learning from 2D Observations

In this scenario, dynamics in the exemplar 3D space are learned using 2D cues only. In that case, the situation is similar when either learning or recognizing. A nice feature here is that only a valid set of 3D exemplars is required, but no additional 3D reconstruction. This is particularly useful when large amounts of 2D observations are available but no 3D inference capabilities (*e.g.* 3D exemplars can be synthesized using a modeling software; the dynamics over these exemplars are learned form real observations).

View observations are not aligned and so the orientation variable $\tilde{l}$ is latent. Nevertheless, the number of latent states remains in practice small, (*i.e.* $L \times N$, with L being the number of discrete orientations $\tilde{l}$ and N the number of states q). The model can be learned by introducing a new variable $\acute{q} = (q, \tilde{l})$ of size $L \times N$ that encodes both state and orientation. Probabilities of this *extended* states are then simply defined as Cartesian products of the transition probabilities for q and $\tilde{l}$. Loops in the model are thus eliminated, and learning can be performed via the forward-backward algorithm introduced in 5.2.1.

6. Action Recognition from 2D Cues

A sequence of observations Y is classified using the MAP estimate (2). Such a probability can now be computed using the classical forward variable $\alpha(\acute{q}_t|\lambda_c) = p(y_1, \ldots, y_t, \acute{q}_t|\lambda_c)$ as explained in [15], where $\acute{q} = (q, \tilde{l})$ is a variable encoding state and orientation as explained in Section 5.2.2

Arbitrary viewpoints do not share similar parameters; in particular scales and metrics can be different. However, the kernel parameter σ^2 is uniquely defined, with the consequence that distances computed in equation (1) can be inconsistent when changing the viewpoint. To adjust σ^2 with respect to changes in these parameters, we introduce $\sigma^2_{\hat{l}} = s_{\tilde{l}}\sigma^2$. Ideally, $\sigma^2_{\hat{l}}$ should be estimated using test data. In practice, the following simple approximation of $\sigma^2_{\hat{l}}$ appears to give satisfactory results with the distance functions we are considering:

$$s_{\hat{l}} = \frac{1}{M}\sum_{i=1}^{M} \frac{\frac{1}{L}\sum_{\tilde{l}=1}^{L} ||P_{\hat{l}\tilde{l}}(\mathsf{x}_i)||^2}{||\mathsf{x}_i||^2}. \qquad (3)$$

Another remark is that observations from multiple calibrated cameras can easily be incorporated. Assuming multiple view observations $\{y_t^1, \ldots, y_t^K\}$ at time t, we can write their joint conditional probability as:

$$p(y_t^1, \ldots, y_t^K|x_t, \hat{l}_t, \tilde{l}_t) \propto \prod_{y_t^k}^{K} p(y_t^k|x_t, \hat{l}_t, \tilde{l}_t). \qquad (4)$$

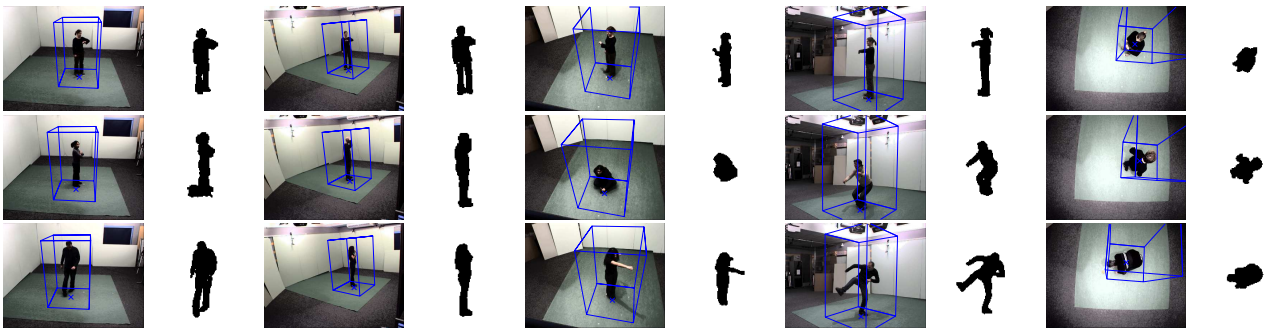

Figure 5. Camera setup and extracted silhouettes: (Top) the action "watch clock" from the 5 different camera views. (Middle and bottom) sample actions: "cross arms", "scratch head", "sit down", "get up", "turn", "walk", "wave", "punch", "kick", and "pick up". Volumetric exemplars are mapped onto the estimated interest regions indicated by blue box.

7. Experiments

Experiments were conducted on our publicly available dataset[1], the IXMAS dataset. We choose 11 actions, performed by 10 actors, each 3 times, and viewed by 5 calibrated cameras (see Figure 5). In this dataset, actor orientations are arbitrary since no specific instruction was given during the acquisition. The 3D sequences are segmented into elementary segments using our approach proposed in [19].

Note, that the same dataset was used in [12] in a similar context. However, results are reported only for a single sequence (out of three) per actor. This sequence has been selected to give best results, thus making a direct comparison difficult.

Our experimental scheme is as follows: 9 of the actors are used for exemplar selection and model learning, the remaining actor is then used for testing. We repeat this procedure by permuting the test-actor and compute the average recognition rate. Examplar selection is performed on subsampled sequences (*i.e.* 2.5 frames/s) to save computational costs. Example results for exemplars are shown in Figure 3. The number M of examplars was empirically set to 52 . Parameter learning and testing is performed using all frames in the database. Action are modeled with 2 states, which appears to be adequate since most segmented actions cover short time periods. Voxel grids are of size: $64 \times 64 \times 64$ and image ROIs: 64×64. The rotation around the vertical axis is discretized into 64 equally spaced values. Consequently, each frame is matched to 52×64 exemplar projections. The ground plane is clustered into 4 positions.

7.1. Learning in 3D

In these experiments, learning is performed in 3D (as explained in 5.2.1). Recognition is then performed on 2D views with arbitrary actor orientations. Recognition rates per camera are given in Figure 6(a), the corresponding views are shown in Figure 5.

cameras	2 4	3 5	1 3 5	1 2 3 5	1 2 3 4
%	81.3	61.6	70.2	75.9	81.3

Table 1. Recognition rates with camera combinations. For comparisons, a full 3D recognition considering 3D manually aligned models as observations, instead of 2D silhouettes, yields 91.11%.

Unsurprisingly, the best recognition rates are obtained with fronto-parallel views (cameras 2 and 4). The top camera (camera 5) scores worst. For this camera, we observe that: the silhouette information is not discriminative; the perspective distortion results in strong bias in distances; estimating the position of the actor is difficult. All these having a strong impact on the recognition performance.

In the next experiment, several views were used in conjunction to test camera combinations. First, 2 view combinations were experimented. Camera 2 and 4 give the best recognition rate at 81.27%. Those 2 cameras are both approximately fronto-parallel and perpendicular one another. Figure 6(b) shows the resulting confusion matrix for this specific setup. Adding further cameras did not improve results. We also try other camera combinations (Table 1). For instance, combining the two cameras with the worst recognition results (camera 3 and 5) raises the recognition rate to 61.59%.

7.2. Learning from single views

In this experiment, learning is performed using single cameras (as explained in Section 5.2.2). Observations during learning and recognition are thus not aligned. The exemplars considered are the same than in the previous section. Learning from a single view is obviously prone to ambiguities, especially when the number of training samples is limited. We thus restricted the experiments to the 3 best cameras with respect to the previous experiments. Figure 6(c) shows the recognition results per action class

[1]The data-set is available on the Perception website http://perception.inrialpes.fr in the "Data" section.

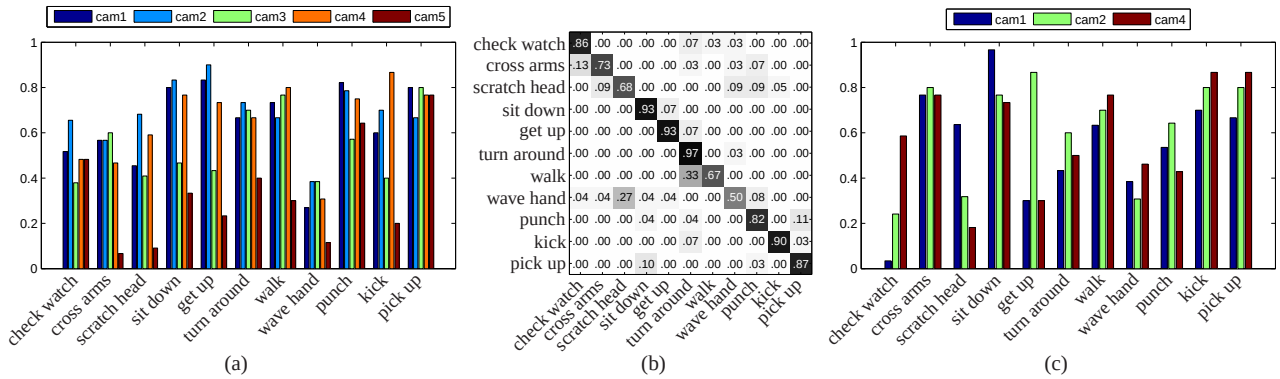

Figure 6. (a) Recognition rates when learning in 3D and recognizing in 2D. The average rates per camera are $\{65.4, 70.0, 54.3, 66.0, 33.6\}$. (b) Confusion matrix for recognition using cameras 2 and 4. Note that actions performed with the hand are confused, *e.g.* "wave" and "scratch head" as well as "walk" and "turn". (c) Recognition rates when learning and recognizing in 2D.

and per camera. Compared to the previous scenario, recognition rates drop drastically, as a consequence of learning from non-aligned data and single view observations. Surprisingly, some of the actions, *e.g.* "cross arms", "kick" still get very acceptable recognition rates, as well as "sit down" and "pick up" that would normally be confused. The average rate for camera 1 is 55.24%, 63.49% for camera 2 and 60.00% for camera 4.

8. Conclusion

This paper presented a new framework for view independent action recognition. The main contribution is a probabilistic 3D exemplar model that can generate arbitrary 2D view observations. It results in a versatile recognition method that adapts to various camera configurations. The approach was evaluated on a dataset of 11 actions and with different challenging scenarios. The best results where obtained with a pair of fronto-parallel perpendicular cameras, validating the fact that actions can be recognized from view arbitrary viewpoints.

References

[1] J. R. Bellegarda and D. Nahamoo. Tied mixture continuous parameter modeling for speech recognition. *ASSP*, 38:2033–2045, 1990. 4

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, pages 1395–1402, 2005. 1

[3] A. Bobick and J. Davis. Real-time recognition of activity using temporal templates. In *WACV*, pages 39–42, 1996. 1

[4] M. Brand. Shadow puppetry. In *ICCV*, pages 1237–1244, 1999. 2

[5] L. W. Campbell, D. A. Becker, A. Azarbayejani, A. F. Bobick, and A. Pentland. Invariant features for 3-d gesture recognition. In *FG*, pages 157–163, 1996. 2

[6] L. W. Campbell and A. F. Bobick. Recognition of human body motion using phase space constraints. In *ICCV*, pages 624–630, 1995. 1

[7] A. A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, pages 726–733, 2003. 1

[8] A. M. Elgammal, V. D. Shet, Y. Yacoob, and L. S. Davis. Learning dynamics for exemplar-based gesture recognition . In *CVPR*, pages 571–578, 2003. 2, 5

[9] B. J. Frey and N. Jojic. Learning graphical models of images, videos and their spatial transformations. In *UAI*, pages 184–191, 2000. 1, 2, 3

[10] D. Gavrila and V. Philomin. Real-time object detection for smart vehicles. In *ICCV*, pages 87–93, 1999. 3

[11] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *ICML*, pages 121–129, 1994. 2, 4

[12] F. Lv and R. Nevatia. Single view human action recognition using key pose matching and viterbi path searching. In *CVPR*, 2007. 2, 3, 4, 6

[13] A. Ogale, A. Karapurkar, G. Guerra-Filho, and Y. Aloimonos. View-invariant identification of pose sequences for action recognition. In *VACE*, 2004. 4

[14] V. Parameswaran and R. Chellappa. View invariance for human action recognition. *IJCV*, 66(1):83–101, 2006. 2

[15] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. pages 267–296, 1990. 2, 5

[16] C. Rao, A. Yilmaz, and M. Shah. View-invariant representation and recognition of actions. *IJCV*, 50(2):203–226, 2002. 2

[17] T. Syeda-Mahmood, M. Vasilescu, and S. Sethi. Recognizing action events from multiple viewpoints. In *EventVideo01*, pages 64–72, 2001. 2

[18] K. Toyama and A. Blake. Probabilistic tracking in a metric space. In *ICCV*, pages 50–59, 2001. 1, 2, 3, 4

[19] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *CVIU*, 104(2-3):249–257, 2006. 1, 2, 6

[20] A. Yilmaz and M. Shah. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *ICCV*, pages 150–157, 2005. 1, 2